学术论文引用预测及影响因素分析*

■ 耿骞 景然 靳健 罗清扬
北京师范大学政府管理学院 北京 100875

摘要:[目的/意义]在引文分析中,可通过论文的一些属性特征对其未来的被引情况进行预测,并通过预测结果对论文、论文作者、作者所属机构及出版物做出评价。[方法/过程]从出版物、作者和论文三个方面对影响论文被引的多个因素展开研究,以图书馆学情报学领域被SCI索引的论文作为分析及验证数据,使用逻辑回归、GBDT、XGBoost、AdaBoost、随机森林等算法进行预测,使用多组评测指标对比不同预测方法的效果,并使用GBDT识别对论文被引影响较大的因素。[结果/结论]确定三个方面的影响因素对论文被引预测的影响程度,构建预测模型,并较好地预测论文在未来一段时间的被引情况。大量实验分析发现GBDT、XGBoost和随机森林的预测能力较强,且预测的时间段越长,效果也就相对越好。

关键词:学术论文 引用预测 影响因素
分类号:G250
DOI:10.13266/j.issn.0252-3116.2018.14.004

1 引言

在科研活动中,学术成果间的引用扮演着重要的角色。研究人员通过引用他人的文章来说明研究背景,阐明学术观点,建立学术研究之间的脉络联系。学术评价工作也常常通过分析论文的被引用情况,间接地评测论文、作者、作者所属机构以及发文期刊的学术影响力。为了度量学术影响力,研究人员已经提出了很多度量指标,其中被引频次是其中最简单、标准和客观的一个度量方法。陈仕吉等[1]提到,在引文分析中,被引频次是用于学术影响力评价的最具代表性的指标。对文献被引频次的讨论一直备受学术界关注,研究人员通过文献的被引频次大小可以识别出重要的学术成果,Google Scholar在对论文排序的时候就把被引频次看作为权重最高的因素[2]。因此,研究人员往往会关注其成果当前及未来的被引用情况,关注影响成果被引的因素,以期提升其成果的被引次数。此外,随着科学研究的发展,每年都有大量新的学术成果发表。同时,由于学科间的交叉融合日趋广泛和深入,很多学术成果会涉及多个研究领域,这样,研究人员很难在有限的时间内关注其研究领域内所有出版物的动态,如果能够预测论文在未来几年内的被引情况,从而间接地确定该论文是否有价值,则能够在一定程度上缓解科研人员搜集和处理论文资料的压力,将时间和精力更多地投入其他的科研活动中。另外,科研管理部门和基金资助机构也希望了解未来哪些成果能获得更多的关注,从而更好地了解学科发展趋势,确定资助领域和课题。

1.1 问题定义

在此前的一些研究中,论文被引预测常被定义为一个回归问题[3-11]。即利用一篇论文的相关特征,来预测这篇论文在未来某时间点的被引频次。虽然这是一种理想的预测方式,然而,在具体进行求解时,为了能够得到较好的预测结果,此类相关研究通常会对数据集合做出一定的预处理以符合实验要求。X. Shi[8]将数据集合中引用次数小于10的论文全部去掉;D. Wang等[9-11]则只使用了在发表前5年内被引频次超过5的论文作为实验数据,这就导致实验数据与真实

* 本文系中央高校基本科研业务费专项资金资助项目"基于社会网络关系的智能专家遴选与推荐平台建设"(项目编号:SKZZB2014037)和教育部人文社会科学研究青年基金项目"面向论文评审专家推荐的兴趣变化挖掘与回避机制生成的研究"(项目编号:16YJC870006)研究成果之一。
作者简介:耿骞(ORCID:0000-0001-5064-4996),教授,博士;景然(ORCID:0000-0002-4191-6221),硕士研究生;靳健(ORCID:0000-0002-3239-2294),副教授,博士,通讯作者,E-mail:jinjian.jay@bnu.edu.cn;罗清扬(ORCID:0000-0003-0750-4028),硕士研究生。
收稿日期:2018-01-09 修回日期:2018-05-22 本文起止页码:29-40 本文责任编辑:徐健

数据分布可能不一致的情况出现；Y. Dong[12] 在论文中也指出，由于论文引用具有明显的长尾效应，因而论文的被引预测其实并不适合采用回归的方式，因而他从另一个角度将被引预测定义为分类问题，即只预测某位作者某篇文章在未来某个时间点文章的被引频次是否能超过作者的 H 指数，如果超过，则说明这篇文章有助于提升作者的影响力，如果没有，则说明这篇文章并没有提升作者的影响力。将被引频次预测从回归问题转变成分类问题以后，由于预测粒度变粗，就可以利用更加符合真实分布的数据，训练出的模型也具有更好的泛化能力，使得研究更有现实价值。因此，在后续的研究中，越来越多的研究[5,13-15] 将论文被引预测定义为分类问题。

综上所述，本研究将论文被引频次预测定义为一个分类预测问题。与 Y. Dong 的研究[12] 不同的是，笔者主要考虑的是作者篇均被引频次而不是 H 指数，主要是因为篇均被引频次相比 H 指数来说更加直观且易于理解。基于此，本研究将论文标记为两类：如果一篇论文在一段时间后获得的被引频次高于作者在论文发表当年的篇均被引频次，则可以说明这篇论文随着不断被引用，对作者影响力的提升起到了一定的正面作用，标记为正类；反之，则标记为负类。本研究的目的是使用多种分类算法和大量论文数据，预测论文在发表一段时间后的被引频次能否超过论文发表当年第一作者的篇均被引频次，即给定论文集合 D，以及发表于时间 t_{d_i} 的某篇论文 $d_i \in D$ 的一系列特征 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ij})$。本研究的任务是训练一个分类模型 C 来预测在时间点 $t_{d_i} + \Delta t$ 时，论文 d_i 的被引频次 $Citation_{d_i,t_{d_i}+\Delta t}$ 能否达到或超过 d_i 的第一作者 $Author_{d_i}$ 在发表当年的平均被引频次 $Citation_{ave,Author_{d_i},t_{d_i}}$，如公式(1)所示：

$$C(d_i|x_i,\Delta t)=\begin{cases}1\ if\ Citation_{d_i,t_{d_i}+\Delta t}\geq Citation_{ave,Author_{d_i},t_{d_i}}\\0\ if\ Citation_{d_i,t_{d_i}+\Delta t}<Citation_{ave,Author_{d_i},t_{d_i}}\end{cases}$$

公式(1)

考虑到论文在发表不同时间后的被引情况不同，对发表不同时间后的论文被引进行预测，预测难度也不同，不同的算法表现也不同，因此在本研究中，将 Δt 分别设定为 1 年、5 年和 10 年，分别代表论文发表初期(发表后不久)，中期(发表一段时间后)和长期(发表很长时间后)，这种选择时间间隔的方式也是在论文引用预测领域被广泛采用的[3-4,16]。

1.2　相关研究

在学术界，人们一直利用被引频次及由其推算出的相关指标衡量学者、研究机构或者研究成果在某一领域的地位。E. Garfield[17] 提出基于被引频次的影响因子来衡量期刊的影响力；J. Hirsch[18] 则提出用 H 指数来衡量学者的影响力。近年来，随着机器学习技术的广泛应用，人们越来越多地关注如何更准确地预测引用情况。M. Callaham 等[19,20] 基于医学类论文，将本领域内的一些特征(如临床分类特征等)加入模型中来进行预测；A. Livne 等[21] 则使用了多个学科的论文数据，最后发现不同学科的预测结果相差较大，在计算机科学、生物学、化学和医学等学科的论文数据中，预测结果表现较好，而在工程、数学和物理论文数据中则表现较差；A. Ibanez[7,22] 等将待预测论文的关键词与高被引论文的关键词间的相关性作为内容特征加入到预测模型中。此外，有研究将社会网络关系加入到预测模型。D. Walker 等[23] 提出了一种基于 PageRank 的方式来预测论文被引频次；刘大有[24] 考虑了论文作者的权威值、引用者的权威值、论文的发表时间以及论文被引用的时间，基于作者和论文间的引用链接，对论文未来被引频次排名和 PageRank 值进行了预测；N. Pobiedina[5] 将论文被引数预测看作是一个链接预测问题，提出一种基于图演化规则的特征 GERscore，并在后续实验中表明 GERscore 可以提升预测的精度；张美平[25] 则结合论文引用的时间衰减特性，提出一种基于持续关注度衰减的重要论文预测算法；F. Davletov[26] 在论文被引预测中，引入论文的拓扑属性(如网络中心度、接近中心度、特征向量中心度等)来改进模型的预测效果。

另外，随着学科交叉现象越来越普遍，不同研究领域之间论文的被引模式可能存在较大的区别。如果直接对某个学科的论文进行引文预测可能会提高预测的难度，降低预测的准确性。因此，有研究人员提出首先对论文的引用模式进行建模，把待被预测的论文归到某个引用模式中，然后再预测论文的被引频次。F. Davletov 等[26] 构建了一个论文的距离矩阵来表征不同的引用模式。然后，他们使用谱聚类方法对所有论文数据进行聚类，随后再通过训练，给每一个类分配一个多项式来进行预测。T. Chakraborty 和 C. T. Li 等[7,27] 基于其定义的一些分类标准，将论文分为几个类别，这样可以将待预测论文分到某一类中后再进行预测。H. Bhat 等[13] 通过估计作者发表成果在学术期刊上的分布，作为作者研究领域分布的近似，用信息熵和 JS 散度来量化每一篇文章的跨学科性，并将跨学科性作为一个新的特征，与一些其他的特征一起来完成预测。

上述研究中虽然使用了各种不同的特征来进行预测，但是对于在使用的众多特征中，哪些特征起到主要作用，哪些特征起到次要作用，不同的特征之间有什么区别等问题并没有较为深入的探讨。另外，在对待引用模式不同的问题时，当前不少相关研究大多是通过提前对不同引用模式进行分类来解决这个问题，但是这种方式的缺点在于，提前对引用模式进行分类使得模型的通用性大大降低，以及他们对引用模式的自定义分类是否科学，目前还尚无定论。因此本研究的创新点在于，笔者首先使用多种机器学习算法对论文未来引用情况进行分类预测，并使用预测中表现较好的一种算法，对论文引用相关因素进行重要性排序，从而甄别出相对重要的因素，并且在预测中引入主题多样性特征和研究方向属性来抓住学科交叉导致的不同引用模式的区别，从而能够构建一种通用的适用于交叉学科的预测模型，提高预测的准确性。

1.3 研究方法与数据集

论文未来的被引趋势受到多种因素的影响。在不同的应用场景下，影响因素的作用强度会不尽相同。因此，确定影响论文被引频次的因素及影响强度是被引预测的核心问题。此外，本研究认为，影响因素的作用不是孤立的，各种因素会综合影响引用行为。准确测量不同因素组合作用的效果对预测有着重要意义。为此，本研究采用相关分析与假设检验的方法，首先对可能影响论文被引次数的因素进行分析，从中挑选出适合引文预测的影响因素类别和因素项，然后使用机器学习算法对论文未来被引情况进行分类预测，并对影响因素在预测中的作用强度进行计算和检测，甄别出影响作用强的因素。考虑到不同的机器学习方法对影响因素作用强度检测效果不同，故笔者采用在分析中具有较好表现的算法进行检测，以从不同角度和方法验证各种因素的影响强度。对采用各种算法时不同类别影响因素组合使用时的性能进行分析，得出不同类别因素对被引预测的影响，对构建不同应用环境下的预测模型提供理论依据。

除了论文本身以外，作者和出版物也是论文撰写和发表过程中两个重要的主体，并且目前已有许多研究人员对被引频次的影响因素进行了研究[19,28-29]，影响因素大概可以分为三类：论文相关因素、作者相关因素和出版物相关因素。因此，本研究也将从作者、出版物和论文三个方面构建影响因素，力求从这三个方面发现影响论文被引趋势的特征。同时，为了能更好地反映当前学科研究的跨学科特点，引入 Web of Science 中论文的研究方向属性特征来更精确地预测论文的被引情况。

用于本研究计算和验证的数据集为 Web of Science 的 Science Citation Index Expanded (SCI-EXPANDED) 数据库中研究方向(研究方向是 Web of Science 下的所有数据库中的论文都使用的一套分类属性，用于对多个数据库中关于同一个主题的文献进行识别、检索和分析)为情报学和图书馆学(Information Science & Library Science)的论文数据，数据包括论文的全信息、参考文献信息和每年的被引频次信息等，时间跨度从 1996 年到 2016 年。另外，本研究还从 Journal Citation Reports 中获取了与试验数据相关的论文所属出版物信息。数据经过简单处理后，共包括 38 442 个作者的 37 677 篇论文。

笔者对数据的来源出版物进行了统计，数据一共来源于46 个出版物，见表 1。从表1 中可以看出，发表研究方向为情报学和图书馆学的论文最多的前 5 名出版物分别是：*Scientist*、*Journal of the American Medical Informatics Association*、*Scientometrics*、*Journal of the American Society for Information Science and Technology* 和 *Information Processing & Management*。

表 1 数据来源出版物统计

来源出版物名称	论文数(篇)
SCIENTIST	7 588
JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	4 503
SCIENTOMETRICS	3 648
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY	2 530
INFORMATION PROCESSING & MANAGEMENT	1 579
ONLINE INFORMATION REVIEW	1 570
INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE	1 471
PROGRAM-ELECTRONIC LIBRARY AND INFORMATION SYSTEMS	1 453
INFORMATION & MANAGEMENT	1 176
JOURNAL OF INFORMATION SCIENCE	1 060

（续表 1）

来源出版物名称	论文数(篇)
SOCIAL SCIENCE COMPUTER REVIEW	1 022
TELECOMMUNICATIONS POLICY	933
EUROPEAN JOURNAL OF INFORMATION SYSTEMS	841
MIS QUARTERLY	813
JOURNAL OF MANAGEMENT INFORMATION SYSTEMS	778
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE	729
INFORMATION TECHNOLOGY AND LIBRARIES	675
JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY	659
JOURNAL OF INFORMATION TECHNOLOGY	612
ASLIB PROCEEDINGS	609
JOURNAL OF STRATEGIC INFORMATION SYSTEMS	448
JOURNAL OF THE ASSOCIATION FOR INFORMATION SYSTEMS	359
ONLINE & CDROM REVIEW	292
ONLINE	256
ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY	194
PROCEEDINGS OF THE ASIS ANNUAL MEETING	163
DATABASE	159
JOURNAL OF DOCUMENTATION	149
RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES	142
JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING	133
ASIST 2001: PROCEEDINGS OF THE 64TH ASIST ANNUAL MEETING, VOL 38, 2001	130
PROGRAM-AUTOMATED LIBRARY AND INFORMATION SYSTEMS	120
ASLIB JOURNAL OF INFORMATION MANAGEMENT	120
DATA BASE FOR ADVANCES IN INFORMATION SYSTEMS	119
ASIST 2003: PROCEEDINGS OF THE 66TH ASIST ANNUAL MEETING, VOL 40, 2003: HUMANIZING INFORMATION TECHNOLOGY: FROM IDEAS TO BITS AND BACK	119
ASIST 2002: PROCEEDINGS OF THE 65TH ASIST ANNUAL MEETING, VOL 39, 2002	116
JOURNAL OF INFORMETRICS	104
BULLETIN OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE	83
DIGITAL LIBRARIES: PEOPLE, KNOWLEDGE, AND TECHNOLOGY, PROCEEDINGS	75
CANADIAN JOURNAL OF INFORMATION AND LIBRARY SCIENCE-REVUE CANADIENNE DES SCIENCES DE L INFORMATION ET DE BIBLIOTH-ECONOMIE	41
HUMAN SOCIETY AND THE INTERNET, PROCEEDINGS: INTERNET-RELATED SOCIO-ECONOMIC ISSUES	36
RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, PROCEEDINGS	26
INFORMATION SYSTEMS JOURNAL	25
VISUAL INTERFACES TO DIGITAL LIBRARIES	17
HYDRODYNAMIC LIMITS OF THE BOLTZMANN EQUATION	1
BEST PRACTICE GUIDELINES ON PUBLISHING ETHICS: A PUBLISHER' S PERSPECTIVE, 2ND EDITION	1
总计	37 677

此外，研究方向为情报学和图书馆学的论文数据体现了当前多学科之间交叉的研究趋势。本研究对其中论文的被引频次进行了统计，将被引频次和计数分别取对数函数作为 X 轴和 Y 轴，见图 1（因许多论文被引频次为 0，无法取对数，所以将所有被引频次加 0.01 后再取对数）。从图 1 中可以发现，该研究方向的论文被引频次和其他很多学科一样，也存在长尾现象：大量论文的被引用次数集中在 0、1 等较低的频次上。这也进一步说明对其进行引用预测并不适合采用回归的方法，而更适合采用分类的方式。

2 影响因素选择

确定影响被引频次因素是进行预测的基础。可以认为，论文的内容是影响其被引次数的重要因素，而出

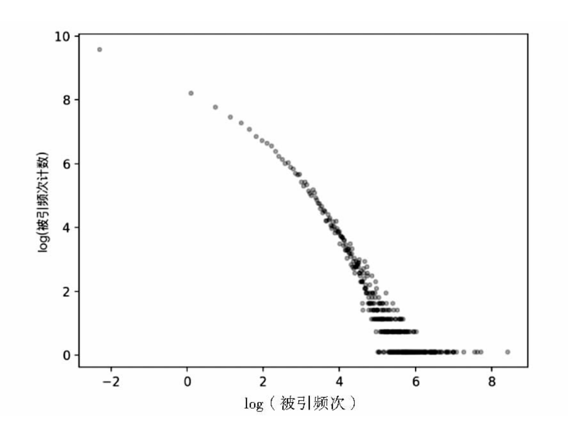


图 1　数据集中论文被引频次的分布

版物和作者的一些特征也在一定程度上影响着论文的被引次数,比如在业内具有高知名度的作者的论文往往更容易被阅读和引用。笔者基于本研究的数据集,提取了能获取到的一些可能影响论文被引次数的影响因素,将影响因素按照其主体分为出版物、作者和论文三大类,使用本文数据集中的数据,将这些影响因素进行计算和统计,从而为后面的预测模型构建和影响因素重要性排序打下基础。表 2 中列出了本文中所使用的所有影响因素名称、来源及其计算和统计的方式。

2.1　出版物相关的影响因素

本研究认为,出版物的质量、水平和学术影响力对

表 2　全部影响因素

类别	影响因素名	影响因素来源与计算统计方式
出版物相关影响因素	总被引数	来源于 JCR 核心指标,实验中使用的数据均为论文出版年时出版物的指标数据。如果个别年份的核心指标数据缺失,则将其余年份数据取平均值对缺失值进行填补;如果出版物在 JCR 中没有被收录,则取其余所有被收录的出版物的核心指标数据的平均值对缺失值进行填补。
	影响因子	
	排除自引后的影响因子	
	五年影响因子	
	即时指数	
	可被引项目数	
	被引半衰期	
	引用半衰期	
	特征因子值	
	论文影响力值	
	可引用项目比	
	标准化特征因子	
	平均影响因子百分位	
作者相关影响因素	第一作者的社会性	建立论文的合作网络,使用公式(2)计算出第一作者在论文出版年时的社会性
	第一作者的 H 指数	统计出在数据集中,论文第一作者在论文出版年之前发表的所有论文,将这些论文按照被引频次从大到小排序,然后根据 H 指数的计算规则,计算出在论文出版年时该作者的 H 指数
	第一作者的论文总数	统计出在数据集中,论文第一作者在论文出版年之前发表的所有论文的数量
	第一作者的过去最大被引数	统计出在数据集中,论文第一作者在论文出版年之前发表的所有论文中,被引最多的论文的被引频次
	第一作者的篇均被引次数	统计出在数据集中,论文第一作者在论文出版年之前发表的所有论文,将这些论文在论文出版年之前的被引频次相加,再除以论文总数,得出篇均被引次数
论文相关影响因素	论文的主题多样性	使用 LDA 计算出每篇论文的主题分布,然后使用公式(4)计算出论文的主题多样性
	论文的页数	来源于 Web of Science 中论文的元数据
	论文参考文献的数量	来源于 Web of Science 中论文的元数据
	论文的研究方向	来源于 Web of Science 中论文的元数据
	论文的使用次数	来源于 Web of Science 中论文的元数据

出版物中所发表的论文的被引频次具有影响作用。出版物本身也具有一些与被引相关的量化指标,包括总被引数、影响因子、排除自引后的影响因子、五年影响因子、即时指数、可被引的文章数、被引半衰期、引用半衰期、特征因子值、论文影响值、被引用的论文数量比、标准化特征因子和平均影响因子百分位等。出版物相关的特征主要来源于 *Journal Citation Reports*,本研究统计出数据集中所有论文的来源出版物,然后在 *Journal Citation Reports* 中可获得出版物的核心指标。

2.2　作者相关的影响因素

作者自身的属性(如学术水平、学术影响力等)和其论文被引有着重要的相关关系。在文献计量中,长期以来对作者的各种属性具有较为全面的研究和指标。本研究尽可能全面地使用已有的相关指标作为其论文被引的影响因素,包括作者的 H 指数、平均被引次数、已发表的论文数、社会性和过去最大被引次数。

其中,H指数能测度科学家以往发表论文的数量和影响力,平均被引次数则为作者的总被引次数除以作者已发表的所有论文数。

作者的社会性与其影响力和被知晓程度具有一定的相关性,因此也和其论文被引具有一定的相关性。作者的社会性越强,其合著者也就越多,其论文被引的机会也就越大。苏芳荔[30]发现合作发表论文的影响力明显高于独立(无合作)发表的论文。社会性的计算方法为:建立一个合作关系网络 $G(A,Co)$,A 是点集,A 中的每一个点 a_i 代表一个作者。Co 是边集,Co 中的每一条边 co_{ij} 代表作者之间的合作关系,边的权重通过合作的论文数来计算。对边的权重进行归一化可得到 a_i 和 a_j 之间的转移概率 $M_{i,j}$,组成转移概率矩阵 M。因此,一个作者 a_i 的社会性 $S(a_i)$ 可通过与其相连的所有其他作者推导出来[4],用公式(2)表示为:

$$S(a_i)=d\sum S(a_j)\cdot M_{j,i}+\frac{1-d}{|A|}\qquad \text{公式(2)}$$

2.3 论文相关的影响因素

论文本身的因素显然应该与论文未来的被引情况直接相关。但是如前所述,由于通过论文内容进行判断的困难性,所以当前的研究都是通过论文的一些方便获得的形式化特征进行推断。

本研究选取的影响因素包括论文的主题多样性、页数、参考文献数量、研究方向属性和使用次数。笔者认为,论文的页数、参考文献数量越多,其内容也就越可能翔实,研究的描述也就越可能细致。此外,论文的研究方向属性越多,说明论文涉及到的研究方向越多,影响面就可能越广,则被引用的可能性就越大。在本文使用的数据集标注了12种研究方向属性,包括计算机科学、情报学与图书馆学、商学与经济学、保健科学与服务、医学信息学、地理学、自然地理学、社会科学－其他、通讯、社会问题、电信、科学与技术－其他。论文的使用次数可以衡量用户对于 Web of Science 平台上一个特定项目的关注程度,该计数反映某篇论文满足用户信息需要的次数,具体表现为用户点击了指向出版商处全文的链接(通过直接链接或 Open URL),或是对论文进行了保存以便在题录管理工具中使用(通过直接导出或保存为之后可以重新导入的其他格式)。论文的使用次数越多,说明论文受到的关注越多,被引用的可能性也就越大。主题(topic)可以理解为特定语料集合下语义的高度抽象和压缩的表示,每一维主题都对应一个比较一致的语义。因此,一篇论文的主题多样性就可以在一定程度下表征该论文研究的多样性程度。笔者利用主题模型中的代表模型隐狄利克雷分布(Latent Dirichlet Allocation,LDA)来计算论文的主题分布。在 LDA 中,主题被表示成 T 个多项式分布,则文档 d 中所有主题的主题分布 $T(d)$ 用公式(3)表示为:

$$T(d)=\{p(topic_1|d),p(topic_2|d)\cdots p(topic_T|d)\}\qquad \text{公式(3)}$$

如果一篇文章的主题具有多样性,那么这篇论文可能被不同研究领域的学者引用,因此被引频次可能会更高。本研究使用信息熵 $D(d)$ 表示论文 d 的主题多样性[4],如公式(4)所示:

$$D(d)=\sum_{i=1}^{T}-p(topic_i|d)\cdot \log p(topic_i|d)\qquad \text{公式(4)}$$

由公式(4)可知,当一篇论文的研究领域较为单一时,则该论文只在某几个主题上有较高的概率分布,其多样性取值较小。当一篇论文涉及多个研究领域时,则这篇论文的主题分布更为均衡,多样性取值也会相对较大。

3 算法选择与模型建立

3.1 预测算法选择

本研究对论文引用进行预测的目的是,将论文引用预测问题定义为一个分类问题,通过数据集训练,发现使用出版物、作者、论文相关影响因素是否能有效地预测出论文在未来的被引频次是否能超过论文发表当年第一作者的篇均被引次数。在分类预测方面,有较多可供选择的方法,本研究选择了朴素贝叶斯(Naive Bayesian Model,NB)、逻辑回归(Logistic Regression,LR)、支持向量机(Support Vector Machine,SVM)、梯度提升决策树(Gradient Boosting Decision Tree,GBDT)、XGBoost(eXtreme Gradient Boosting)、AdaBoost(Adaptive Boosting)和随机森林(Random Forest,RF)7种算法。使用这些算法来进行预测的原因是,朴素贝叶斯、逻辑回归和支持向量机是三种经典的分类算法,并且已经在不同的数据集合中证明其有效性,而 GBDT、XGBoost、AdaBoost 和随机森林都属于效果较为优异的集成学习算法,特别是 GBDT、XGBoost、AdaBoost 等集成学习算法由于其出色的泛化能力在近几年被广泛应用于学术研究和实际工作中。

3.2 评测指标选择

本研究使用分类器性能评价常用的指标 ROC(Receiver Operating Characteristic Curve)曲线下面积(Area Under Curve,AUC)和 F1 值来进行评测,这两个指标也常常被论文引用分类预测领域用来对预测效果进行评测[12-13,31]。要计算 AUC 和 F1 值,首先需要对

一些指标进行定义,如果实际值属于正类,预测值也为正类,则标记为 TP;如果实际值属于正类,预测值为负类,则标记为 FN;如果实际值为负类,预测值为正类,则标记为 FP;如果实际值为负类,预测值也为负类,则标记为 TN,如表 3 所示:

表 3　分类结果的混淆矩阵

实际值 \ 预测值	正	负
正	TP	FN
负	FP	TN

(1)AUC。AUC 和 ROC 常被用来评价一个二分类器的分类性能,AUC 值越大,说明分类器的效果越好。ROC 曲线坐标轴的横坐标叫做"假正例率",用符号表示为 FPR,纵坐标叫做"真正例率",用符号表示为 TPR。FPR 和 TPR 的计算方法如下所示:

$$FPR = \frac{FP}{TN + FN} \qquad \text{公式(5)}$$

$$TPR = \frac{TP}{TP + FP} \qquad \text{公式(6)}$$

(2)F1 值。F1 值是准确率(precision)和召回率(recall)的调和均值。准确率是预测出的正样本中,真实正样本的比例;召回率是真实的正样本中,被正确预测为正样本的比例。准确率和召回率越高,说明模型效果越好。但是准确率和召回率常常是相互制约的,因此 F1 值用来对准确率和召回率进行加权调和,其公式如下所示:

$$F1\text{ 值} = \frac{2 \cdot precision \cdot recall}{precision + recall} \qquad \text{公式(7)}$$

3.3　影响因素抽取

(1)主题多样性计算。首先利用 NLTK(Natural Language Toolkit)对数据集中论文的题名和摘要进行分词和词性标注。NLTK 是 Python 环境中用于自然语言处理的工具包。本研究仅保留名词性短语和形容词性短语,并利用混杂度作为选取主题数的准则,超参数 α 设置为 0.01,β 设置为 $50/k$(k 为主题个数),得到每一篇论文的主题分布 $T(d)$。根据公式(4),计算每一篇论文的主题多样性 $D(d)$。

(2)作者社会性计算。建立合作关系网络 $G(A, Co)$,参数 d 取 0.85。使用公式(2)计算出作者的社会性 $S(a)$。

(3)对离散型变量进行 one-hot 编码。对每一篇论文的研究方向属性进行 one-hot 编码,因为数据集中共有 12 种研究方向属性,因此将每一篇论文的研究方向属性转换为一个 12 维的二进制向量。论文属于哪个研究方向属性,就将该研究方向属性对应的维度标记为 1,其余标记为 0。

(4)计算和统计其他影响因素。对除了主题多样性、作者社会性、研究方向属性以外的因素进行计算和统计,比如作者的 H 指数、平均被引次数、已发表的论文数和过去最大被引次数等,这些影响因素的计算和统计都是基于本文中所采用的数据集。

3.4　数据预处理

(1)数据标注。基于本文的问题定义,设定 Δt 分别为 1 年、5 年和 10 年,把论文在出版 Δt 后的真实被引频次 $Citation_{d_i, t_{d_i}+\Delta t}$ 与第一作者 $Author_{d_i}$ 在出版当年的篇均被引频次 $Citation_{ave, Author_{d_i}, t_{d_i}}$ 相比较。如果被引频次大于作者在出版当年的篇均被引频次,则认为此论文对提升作者的影响力起正面作用,标注为正类,如果被引频次小于作者在出版当年的平均被引频次,则认为此论文没有起到提升作者影响力的作用,标注为负类。

(2)归一化影响因素。将抽取出来的所有影响因素进行归一化处理,把每一维度的原始数据都等比例缩放到[0,1]范围内。归一化公式为:$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$,其中 X_{norm} 为归一化后的数据,X 为原始数据,X_{max} 和 X_{min} 分别为原始数据集中的最大值和最小值。

3.5　预测建模

将 70% 的数据作为训练集,30% 的数据作为测试集,分别使用所有影响因素、单类影响因素和两两组合影响因素,使用上文所述 7 种方法来建立预测建模。模型的训练过程如图 2 所示:

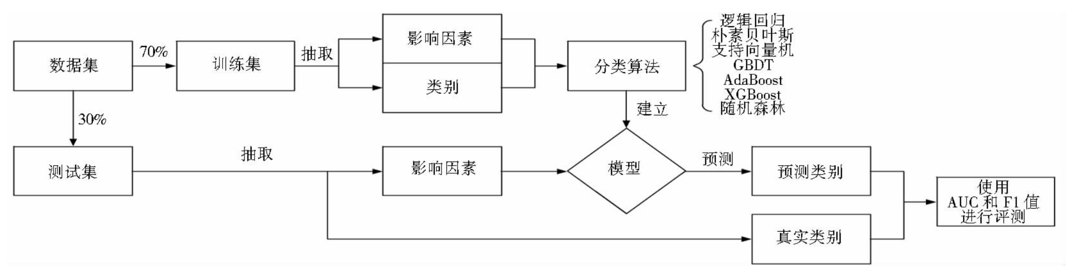


图 2　模型训练过程

首先抽取数据集中的影响因素，将数据类别进行标注，其中的 70% 数据作为训练集，30% 数据作为测试集；然后将训练集中的影响因素和类别使用分类算法进行训练建立分类模型，然后将测试集中的影响因素输入分类模型，分类模型输出其预测的类别。这里选择其中一种分类算法 GBDT 来介绍模型的训练过程，其流程如图 3 所示：

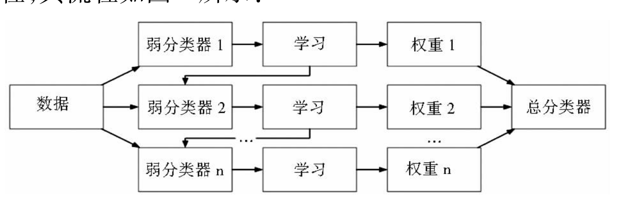


图 3 GBDT 训练过程

GBDT 通过多轮迭代，每一轮迭代都产生一个弱分类器（一般使用 CART 作为弱分类器），使用上一个分类器的残差训练出下一个弱分类器，最后将每一轮迭代产生的弱分类器加权求和，得到总分类器。

本研究使用 python 的 scikit-learn 训练 GBDT，scikit-learn 封装了 GBDT 的类库，其中 GradientBoostingClassifier 是用于分类的类，GradientBoostingClassifier 的参数分为两类：一类是 Boosting 框架的参数，另一类是弱学习器的参数。

Boosting 框架的重要参数包括：①*n_estimators*，设置了弱学习器的最大迭代次数，也就是弱学习器的最大个数，该参数太大模型会过拟合，太小则会欠拟合，默认是 100；②*learning_rate*，设置的是每个弱学习器的步长，默认是 1，步长越小，迭代次数就越多，因此 *n_estimators* 和 *learning_rate* 常常一起调整；③*loss*，也就是损失函数，分类模型有两种损失函数，对数似然损失函数"deviance"和指数损失函数"exponential"，GBDT 使用的是对数似然损失函数"deviance"。

弱学习器的重要参数包括：①*max_features*：表示在划分的时候考虑的最大特征数，如果特征较多，则需要设置该参数，以减少训练时间，默认值是"None"；②*max_depth*：表示作为弱分类器的决策树的最大深度，与 *max_features* 类似，如果特征较多，则需要降低最大深度以减少训练时间，默认不输入；③*min_samples_split*：表示内部节点再划分所需要的最小样本数，默认值为 2，如果某节点的样本数少于该参数设置的数，则不再进行划分；④*min_samples_leaf*：叶子节点最少样本数，如果某叶子节点的数目小于这个参数设置的数字，则会和兄弟节点一起被剪枝，默认值为 1。

最后，将分类模型预测出的类别与真实类别进行比较，计算出模型的 AUC 和 F1 值，对模型的效果进行评价。

4 实验结果与分析

4.1 论文引用预测

4.1.1 使用所有影响因素时的算法比较 首先使用所有影响因素和不同算法进行实验，将 7 种算法实验结果的 AUC 和 F1 值在柱状图中呈现。其中，X 轴是 Δt，$\Delta t \in \{1$ 年，5 年，10 年$\}$，Y 轴是评测指标的大小，分别如图 4 和图 5 所示：

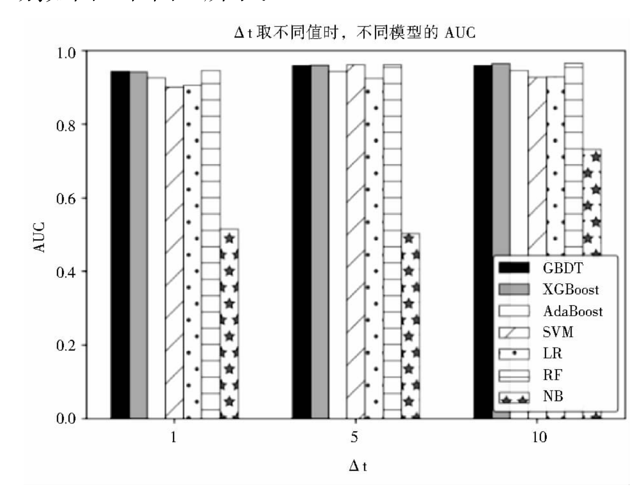


图 4 Δt 取不同值时，不同模型的 AUC

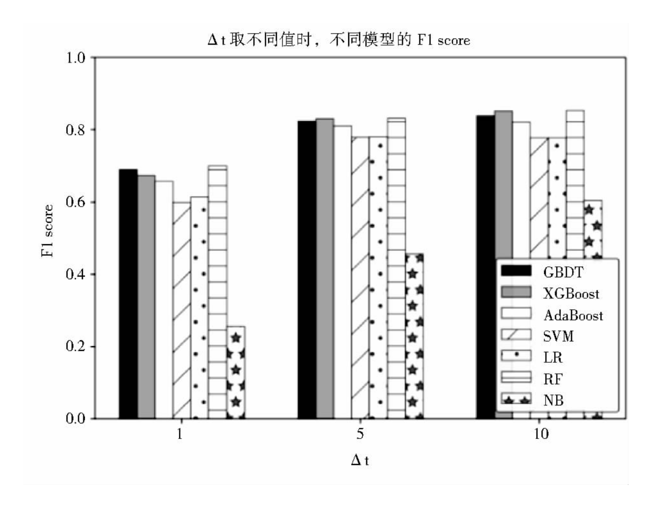


图 5 Δt 取不同值时，不同模型的 F1 值

从图中可以看出，在 Δt 分别为 1 年、5 年和 10 年时，GBDT、XGBoost 和随机森林在 AUC 和 F1 值指标上取得了最好的结果。其中，XGBoost 和随机森林在 F1 值和 AUC 上分别达到了 0.85 和 0.96 以上的分数。该结果证明当前的影响因素选择方式和算法选择对于论文被引预测是有效的，也证明集成学习算法适用于论文引用预测这一领域。

4.1.2 不同类别影响因素对预测的影响 为了进一步检验作者、出版物和论文三个类别的影响因素在预测中的作用,本研究又分别对这三类影响因素单独进行了试验。然后,将这三类影响因素两两组合进行试验,分别考察不同情况下的表现。考虑到当前预测 10 年的效果最好,因此只选择时间间隔为 10 年。选择表现最好的三个算法 GBDT、XGBoost 和随机森林进行试验,结果如图 6－图 9 所示:

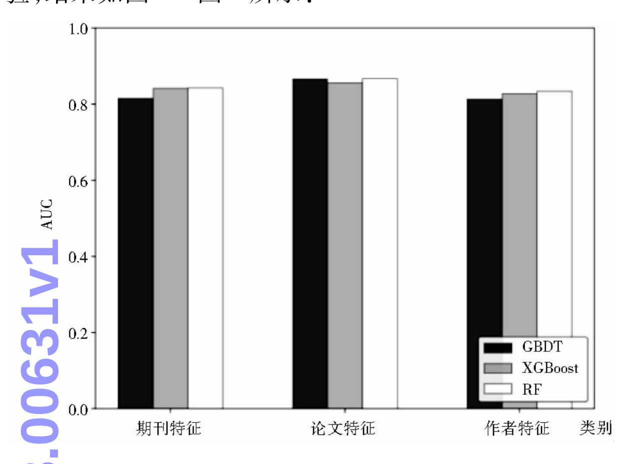


图 6 使用单类影响因素时,不同模型的 AUC

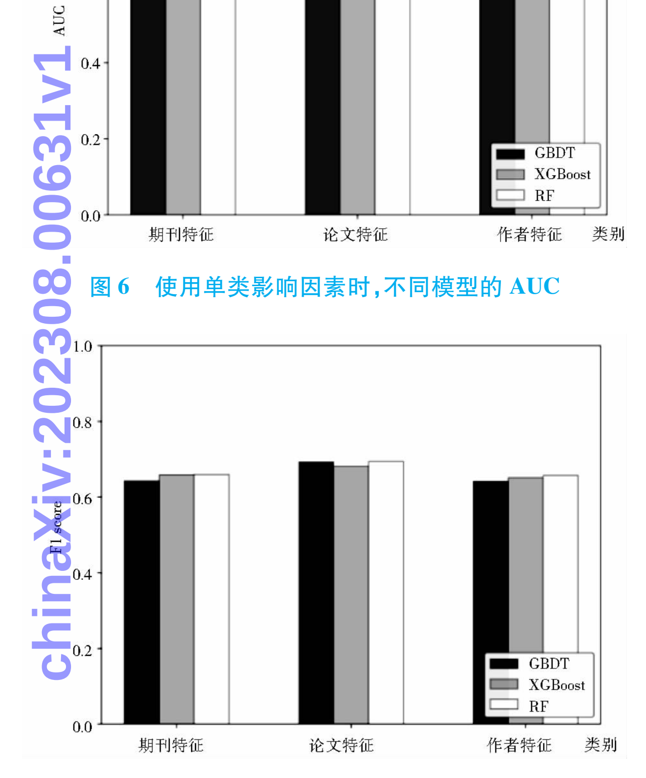


图 7 使用单类影响因素时,不同模型的 F1

从结果中可以看出,单独使用某一类影响因素的效果都要逊色于使用全部影响因素的效果。其中,只利用论文相关影响因素的效果要略好于其他两类影响因素,而影响因素两两组合后的效果要优于只采用单一某类影响因素,但是仍然逊色于使用全部影响因素的预测效果。其中,“作者＋论文”的影响因素组合是两两组合中效果最好的。综上结果可以看出任一类下的影响因素或者影响因素的两两组合效果都不如全特征下的效果好。即在本数据展开的预测中,利用的特征越多,预测也就越准确。

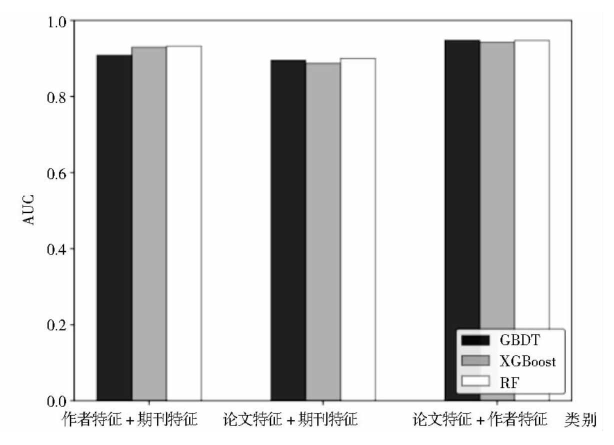


图 8 影响因素两两组合时,不同模型的 AUC

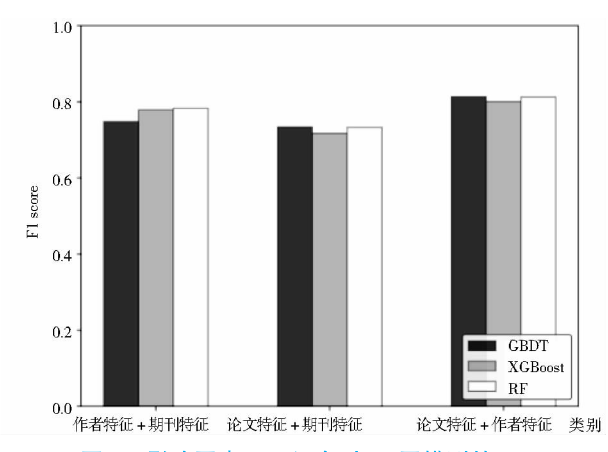


图 9 影响因素两两组合时,不同模型的 F1

4.2 影响因素重要性排序

在进行了论文引用预测后,本研究使用了梯度提升决策树(Gradient Boosting Decision Tree,GBDT)对影响因素的重要性进行了排序。GBDT 是由多棵决策树迭代组成,每一颗树迭代的过程中都会做特征选择,通过特定的衡量指标,从候选特征中选择一个特征及相应的分裂值,特征所处的树层次越接近根节点,分裂次数越多,特征就越重要,特征 j 的重要性计算方法为由 J. H. Friedman[32] 提出,计算公式如下所示:

$$\hat{I}_j^2(T)=\sum_{t=1}^{J-1}\hat{i}_t^2(v_t=j) \qquad 公式(8)$$

其中,树 T 有 J 个叶子节点,则非叶子节点有 $J-1$ 个,v_t 是跟节点 t 相关的分裂特征,$\hat{i}_t^2$ 是对应节点 t 分裂后减少的平方损失。而对于包含了 M 棵树的森林 $\{T_m\}_1^M$ 来说,特征 j 的全局重要性可以通过其在所有树上的重要性平均值推导出来,计算公式如下所示:

$$\hat{I}_j^2=\frac{1}{M}\sum_{m=1}^{M}\hat{I}_j^2(T_m) \qquad 公式(9)$$

在本实验中,可以利用 GBDT 输出所有影响因素的重要性分布,表 4 列出了在时间间隔取不同的值时,排名前 10 位的影响因素的重要性。

表 4 影响因素重要性 TOP10

时间间隔	$\Delta t=1$		$\Delta t=5$		$\Delta t=10$	
排名	影响因素	重要性	影响因素	重要性	影响因素	重要性
1	第一作者的平均被引数	0.328 2	论文的使用次数	0.260 8	论文的使用次数	0.224 1
2	论文的使用次数	0.161 5	第一作者的平均被引数	0.253	论文页数	0.219 2
3	第一作者的最大被引数	0.112 1	参考文献数量	0.132 8	第一作者的平均被引数	0.182 1
4	参考文献数量	0.107	论文页数	0.108 9	参考文献数量	0.134 5
5	总被引数	0.048 48	第一作者的最大被引数	0.107 6	第一作者的最大被引数	0.105 8
6	即时指数	0.035 53	论文的主题多样性	0.019 98	第一作者的论文总数	0.015 67
7	排除自引后的影响因子	0.034 5	总被引数	0.018 89	论文的主题多样性	0.014 26
8	论文页数	0.029 53	排除自引后的影响因子	0.016 77	总被引数	0.014 01
9	第一作者的论文总数	0.025 99	第一作者的论文总数	0.010 26	被引半衰期	0.012 65
10	论文的主题多样性	0.018 59	即时指数	0.010 08	排除自引后的影响因子	0.012 58

从表 4 中可以看出,在时间间隔分别取 1、5 和 10 年时,论文相关的影响因素和出版物相关的影响因素在 GBDT 的训练中都起到了较为重要的作用。其中论文相关的影响因素中,论文的使用次数和参考文献数量在三个时间间隔中都排在了较前的位置,这表明论文被浏览下载的次数越多,越有可能被引用,论文的参考文献越丰富,论文可能前期的调研工作越扎实。另外,在作者相关的影响因素中排名靠前的是第一作者的平均被引次数和第一作者的最大被引次数。一般而言第一作者是论文的撰写者,是直接决定论文内容的人,因此第一作者自身的学术水平会较大地影响论文是否会被潜在的引用。

在对所有影响因素进行了重要性排序后,本研究利用 GBDT 分析了使用单类影响因素时的重要性,表 5 列出了使用单类影响因素时输出重要性前 5 位的影响因素。

表 5 单类影响因素重要性排序 TOPS

单类影响因素重要性 TOP5					
出版物影响因素		论文影响因素		作者影响因素	
总被引数	0.254 6	论文页数	0.422 2	第一作者的平均被引数	0.36
影响因子	0.155 5	论文使用次数	0.304 6	第一作者的最大被引数	0.284 6
即时指数	0.135 1	参考文献数量	0.168 2	第一作者的社会性	0.179 5
引用半衰期	0.101 5	论文的主题多样性	0.051 59	第一作者的论文总数	0.141 5
排除自引后的影响因子	0.085 8	论文的研究方向	0.008 749	第一作者的 H 指数	0.034 2

另外,从特征重要性表中可以发现,单一大类影响因素与全部影响因素下的特征排序大致是相同的。比如在单一大类影响因素中,在论文相关影响因素中排名第二位的论文使用次数和排名第三位的论文参考文献数量在全部影响因素排序下是属于论文相关因素的前两位。在作者相关影响因素排序中,排名第一的第一作者平均被引次数和排名第二的第一作者最大被引次数在全部影响因素排序下也属于作者相关影响因素的前两位。这在一定程度上相互验证了全部影响因素排序下输出的重要性排序和单类影响因素重要性排序的正确与否。

5 结论与后续研究

本研究对与引用预测有关的影响因素进行了梳理分类,得到作者、出版物和论文三类影响因素,选取了 Information Science & Library Science 学科进行试验,首次梳理出论文引用预测的影响因素重要性排序,并且在实验过程中,引入了 GBDT、XGBoost、AdaBoost 等一系列集成学习方法进行预测,取得了较好的效果。

从研究结果中可以看出:

(1)当 Δt 分别取 1 年、5 年和 10 年时,随着 Δt 的增大,7 种算法的预测能力都有明显的提升,说明时间间隔越长,论文的被引情况就越趋于稳定,预测的效果也就越好;

(2)在 7 种算法中,本研究所引入的集成学习算法,如 GBDT、XGBoost 和随机森林取得了最好的预测效果,说明集成学习算法能很好地应用于论文引用预测中;

(3)通过影响因素重要性排序分析发现,作者相关的影响因素和论文相关的影响因素比出版物相关的

影响因素对论文引用预测的影响更大。在作者相关的影响因素中,作者的篇均被引数和最大被引数的重要性较高,说明作者的被引数在一定程度上代表了作者在其研究领域中的影响力,被引数高的作者能吸引到更多的引用;在论文相关影响因素中,论文的使用次数、参考文献数量和页数相对论文的内容特征来说更为重要。这也与人们的日常认知相符,参考文献数量和页数表征了作者前期调研和后期研究的扎实程度,而论文的使用次数则反映了论文的受欢迎程度,使用次数多,就能吸引更多的引用;而在出版物相关影响因素中,相对于被业界广为认可的影响因子来说,排除自引后的影响因子反而体现出了更强的重要性,说明自引对于提升学术影响力并没有什么太大的作用。

本研究目前仅基于已有的研究工作,将论文的引用预测定义为二分类问题,在后续研究中,可以对问题进行更为细致的定义,比如使用更加细粒度的分类。另外,将采用覆盖更多学科的数据集合,针对更多的学科数据进行研究,提取出更多可能影响论文被引的因素,以期形成一个较为完整的论文引用预测方法框架。

参考文献:

[1] 陈仕吉，史丽文，左文革. 基于 ESI 的学术影响力指标测度方法与实证[J]. 图书情报工作, 2013, 57(2): 97－102,123.

[2] BEEL J, GIPP B. Google Scholar's ranking algorithm: the impact of citation counts (an empirical study)[C]// Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science. Piscataway: IEEE, 2009: 439－446.

[3] YAN R, TANG J, LIU X, et al. Citation count prediction: learning to estimate future citations for literature[C]// Proceedings of the 20th ACM Conference on Information and Knowledge Management. Glasgow: ACM, 2011: 1247－1252.

[4] YAN R, HUANG C, TANG J, et al. To better stand on the shoulder of giants[C]// Proceeding of the ACM/IEEE Joint Conference on Digital Libraries. Washington: ACM, 2012: 51－60.

[5] POBIEDINA N, ICHISE R. Predicting citation counts for academic literature using graph pattern mining[C]// International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems. Kaohsiung: Springer, 2014: 109－119.

[6] POBIEDINA N, ICHISE R. Citation count prediction as a link prediction problem[J]. Applied intelligence, 2016, 44(2): 252－268.

[7] CHAKRABORTY T, KUMAR S, GOYAL P, et al. Towards a stratified learning approach to predict future citation counts[C]// Proceeding of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. London: Piscataway, 2014: 351－360.

[8] SHI X, LESKOVEC J, MCFARLAND D A. Citing for high impact [C]// Proceeding of the 10th annual joint conference on Digital libraries. Queenland: ACM, 2010: 49－58.

[9] WANG D, SONG C, BARABÁSI A L. Quantifying long-term scientific impact[J]. Science, 2013, 342(6154): 127－132.

[10] SHEN H W, WANG D, SONG C, et al. Modeling and predicting popularity dynamics via reinforced poisson processes[C]//Proceeding of the 28th AAAI Conference on Artificial Intelligence. Quebec: AAAI, 2014: 291－297.

[11] XIAO S, YAN J, LI C, et al. On modeling and predicting individual paper citation count over time[C]// Proceeding of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: Morgan Kaufmann, 2016.

[12] DONG Y, JOHNSON R A, CHAWLA N V. Will this paper increase your h-index?: Scientific impact prediction[C]// Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2015. Porto: Springer, 2015: 149－158.

[13] BHAT H S, HUANG L H, RODRIGUEZ S, et al. Citation prediction using diverse features[C]// 2015 IEEE International Conference on Data Mining Workshop. Atlantic: IEEE, 2015: 589－596.

[14] MCKEOWN K, DAUME H, CHATURVEDI S, et al. Predicting the impact of scientific concepts using full-text features[J]. Journal of the association for information science and technology, 2016, 67(11): 2684－2696.

[15] NEZHADBIGLARI M, GONÇALVES M A, ALMEIDA J M. Early prediction of scholar popularity[C]// Proceeding of the ACM/IEEE Joint Conference on Digital Libraries. Newark: IEEE, 2016: 181－190.

[16] ACUNA D E, ALLESINA S, KORDING K P. Future impact: predicting scientific success[J]. Nature, 2012, 489(7415): 201－202.

[17] GARFIELD E. Citation indexes for science: A new dimension in documentation through association of ideas[J]. Science, 1955, 122(3159): 108－111.

[18] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the national academy of sciences, 2005, 102(46): 16569－16572.

[19] CALLAHAM M, WEARS R L, WEBER E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals[J]. Jama, 2002, 287(21): 2847－2850.

[20] KULKARNI A V, BUSSE J W, SHAMS I. Characteristics associated with citation rate of the medical literature[J]. PloS one, 2007, 2(5): e403.

[21] LIVNE A, ADAR E, TEEVAN J, et al. Predicting citation counts using text and graph mining[C]// iConference. Fort Worth: Morgan & Claypool Publishers, 2013.

[22] IBÁÑEZ A, LARRAÑAGA P, BIELZA C. Predicting citation count of Bioinformatics papers within four years of publication[J]. Bioinformatics, 2009, 25(24): 3303－3309.

[23] WALKER D, XIE H, YAN K K, et al. Ranking scientific publications using a model of network traffic[J]. Journal of statistical mechanics: theory and experiment, 2007 (6): P06010.
[24] 刘大有, 齐红, 薛锐青. 基于作者权威值的论文价值预测算法[J]. 自动化学报, 2012, 38(10): 1654 – 1662.
[25] 张美平, 尚明生. 基于持续关注度衰减的重要论文预测[J]. 复杂系统与复杂性科学, 青岛大学, 2015, 12(3): 77 – 84.
[26] DAVLETOV F, AYDIN A S, CAKMAK A. High impact academic paper prediction using temporal and topological features[C]// Proceeding of the 23rd ACM international conference on information and knowledge management. Shanghai: ACM, 2014: 491 – 498.
[27] LI C T, LIN Y J, YAN R, et al. Trend-based citation count prediction for research articles[C]// Advances in knowledge discovery and data mining. Ho Chi Minh: Springer, 2015: 659 – 671.
[28] BUELA-CASAL G, ZYCH I. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals[J]. Psicothema, 2010, 22(2): 270 – 276.
[29] JABBOUR C J C, JABBOUR A B L de S, DE OLIVEIRA J H C. The perception of brazilian researchers concerning the factors that influence the citation of their articles: A study in the field of sustainability[J]. Serials review, 2013, 39(2): 93 – 96.
[30] 苏芳荔. 科研合作对期刊论文被引频次的影响[J]. 图书情报工作, 2011, 55(10): 144 – 148.
[31] DONG Y, JOHNSON R A, CHAWLA N V. Can scientific impact be predicted? [J]. IEEE transactions on big data, 2016, 2(1): 18 – 30.
[32] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001, 29(5): 1189 – 1232.

作者贡献说明:
耿骞:研究设计、论文大纲和论文修改;
景然:研究设计与实现、论文撰写;
靳健:研究方案设计,论文修改;
罗清扬:文献调研,数据整理。

Citation Prediction and Influencing Factors Analysis on Academic Papers

Geng Qian Jing Ran Jin Jian Luo Qingyang
School of Government, Beijing Normal University, Beijing 100875

Abstract: [**Purpose/significance**] In this study, the prediction about future citation of a paper is analyzed by a set of features, which intends to evaluate the academic influence of a scholar, a paper and/or a publication. [**Method/process**] In this study, publications, authors and papers are investigated to discuss potential factors for citation prediction and SCI indexed papers in the field of Library Information are utilized as a concrete example to evaluate the validity of these factors. Several algorithms, such as logistic regression, GBDT, XGBoost, AdaBoost and Random Forest, are benchmarked on different evaluation metrics and the algorithm of GBDT is applied to identify influential factors. [**Result/conclusion**] Three aspects of influential factors for citation prediction are analyzed and different approaches are evaluated, which aims to predict citations of papers in the near future. Categories of experiments are conducted and it is found that GBDT, XGBoost and Random Forest perform the best. Also, the performance of citation prediction tends to be better on papers with a relative longer publication time.

Keywords: academic papers citation prediction influencing factors